MAMMOGRAPHY IMAGE CLASSIFICATION USING TEXTURE FEATURES

Blagojce Jankulovski, Ivan Kitanovski, Katarina Trojacanec, Ivica Dimitrovski, Suzana Loskovska
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

ABSTRACT

Mammography image classification is a very important research field due to its implementation domain. The aim of this paper is propose techniques for automation of the mammography image classification process. This requires the images to be described using feature extraction algorithms and then classified using machine learning algorithms. In that context, the goal is to find which combination of feature extraction algorithm and classification algorithm yield the best results for mammography image classification. The following feature extraction methods were used LBP, GLDM, GLRLM, Haralick, Gabor filters and a combined descriptor. The images were classified using several machine learning algorithms i.e. support vector machines, random forests and k-nearest neighbour classifier. The best results were obtained when the images were described using GLDM together with the support vector machines as a classification technique.

I. INTRODUCTION

One of the most common causes for women mortality all over the world is considered to be the breast cancer. The cause of the cancer are poorly understood, thus medical experts still cannot find a way to prevent it. Chances of survival are increased if the cancer is detected in its early stage. There are three methods used for breast cancer detection: mammography, biopsy and needle aspirate. The first step towards breast cancer detection is mammography. Mammography is a process where a patient's breast is exposed to low does x-rays thus producing a mammographic image. It is most frequently used because it is a cheap and non-invasive method for investigating patient's health status and allows early stage cancer detection. The mammographic image is usually analyzed by doctors or radiologists and if they find cancerous or suspicious regions, they send the patient to additional test, such as biopsy and needle aspirate [1]. Manual analysis depends of many factors including the experience of the medical practitioner or radiologist. Furthermore, this is a repetitive task which requires a lot of attention to minor details. Reports show that cancer detection rate using mammography is between 70-90% [2]. This means that around 10-30% of breast cancers are missed in this early stage, which can result in severe future problems.

In an attempt to improve early detection, our research is focused on the mammograms analysis procedure aiming to improve the differentiation of benign from malignant cases, or detect suspicious cases in general. Therefore we compare several techniques for automated mammography image classification.

The paper is organized as follows. Section 2 briefly describes the problem and the motivation to our approach. Section 3 presents the feature extraction techniques used for describing the visual content of the images. Section 4 presents the classification methods used in the paper. The experimental setup is thoroughly explained in section 5. The experimental results and a brief discussion of them are provided in section 6. Section 7 gives the concluding remarks of this work.

II. PROBLEM DESCRIPTION

Automated mammography image classification refers to the process of unsupervised labelling of images as normal or abnormal i.e. containing suspicious tissue. It is the process of determining whether mammographic images contain certain abnormalities or not.

Image classification can be performed directly, through direct classification of the image by pixel value. The more common way to classify an image is by using feature extraction algorithms to describe the image visual content. The feature extraction algorithms generate descriptions which are called descriptors. Then, classification is performed on those descriptors. Sometimes there is a pre-processing stage prior to feature extraction [3], which is usually done to improve image contrast, to segment regions of interest and to do other tuning operations, which helps to extract more appropriate features.

Automated classification of mammograms is an area of active research [3]. Most of the work in this field is focused on detection and classification of micro-calcifications, circumscribed masses and speculated lesions [4][5][6]. Others direct their research towards classifying lesions as benign or malign [7]. In [8], the images are first pre-processed so that the image content is enhanced for better feature extraction. The classification phase is performed using C4.5 algorithm. In [9] the authors first extract the image content using a bag-of-words method and then use k-nearest neighbours for the classification process. In [10] Haralick features are used for describing the visual content of the images and Bayesian Neural Network classifier is used for the classification phase.

Most of the work research in this field is by using one feature extraction method and one classification algorithm. Our research focuses on experimenting with different types of feature extraction algorithms and classification methods to compare and find the best suited combination.

It is obvious that most of the work done in this field is by using texture based features extraction algorithms to describe the image content. If we analyze the nature of mammographic images, we can see that they are greyscale images containing a series of patterns. Colour descriptors would have no effect here, since the images are greyscale. Shape descriptors can be used as well, but given the nature of the images and the proven and efficient way of describing textures we also use texture based descriptors for describing mammographic images.

In the classification phase there are various approaches. From [8], it is obvious that C4.5 alone is not appropriate for this type of data. We propose the use of random forest classifier, which is among the most efficient classifiers today, support vector machines and k-nearest neighbour classifier which performed well in [9].

III. Feature extraction

The first step in the classification process is extracting the visual features of the mammographic images. The main problem facing this process is the different resolution, quality and the week contrast of the mammograms. This makes the detection of the cancer much harder. An example is shown on Fig. 1.a. To overcome this problem and make the feature extraction as efficient as possible, pre-processing of the image is required. We used the histogram equalization to adjust the image contrast. The method improves the contrast by spreading out the most frequent intensity values. It is useful in images with backgrounds and foregrounds that are both bright and both dark. Histogram equalization is used to make contrast adjustment so the anomalies can be better emphasized, as shown on Fig.1b.

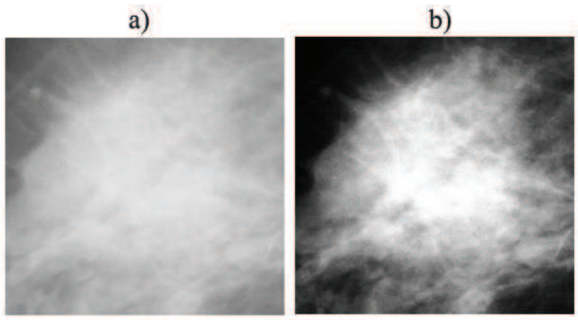

Figure 1. Mammographic image of left and right breast

We implemented six texture descriptors, LBP (Local Binary Patterns), GLDM (Gray Level Difference Method), GLRLM (Grey Level Run Length Method), Haralick texture features, Gabor texture filters and combined descriptor.

Local Binary Patterns is a method that uses gray-scale invariant texture statistics. LBP associates a binary value to each pixel on the bases of the similarity between that pixel and his neighbour pixels [13][14]. We implemented the LBP feature descriptor with 59 features.

We also use the gray level difference method (GLDM) [11], [12]. Four possible forms of the vector δ are considered, (0, d), (-d, d), (d, 0), (-d, -d), where d is the inter sample spacing distance. For different values of d we calculated five texture features: Contrast, Angular Second Moment, Entropy, Mean and Inverse Difference Moment. The GLDM feature descriptor is calculated for five displacements (1, 2, 3, 4 and 5) and, thus, the implementation has 25 features.

The GLRLM method is based on calculating the number of gray level runs of various lengths [11], [15]. The length of the run is defined as the number of consecutive pixels having the same gray level value. For each direction, gray level run length matrices are computed, which are then used to compute five features for each matrix. We used the 44 run feature descriptor.

Haralick texture feature is based on the co-occurrence matrix used for displaying the gray level spatial dependency different angular relationships, vertical and horizontal directions in the image. Usually this feature descriptor has 13 features calculated from the co-occurrence matrix [16].

Gabor texture filter is a linear filter used for edge detection [17]. Gabor filter which contains 48 features is used for this research.

The Combined descriptor is basically a combination of the previous 5 descriptors. The feature vector is constructed by merging the feature vectors of all other descriptors into a single vector.

IV. Classification of mammograms

There are various algorithms for automated classification. We used several classification algorithms to compare their performance: support vector machines, random forests and k-nearest neighbour classifier.

Support vector machines (SVM) are among the most powerful classifiers today [18][19][20]. Their primary goal was to solve binary classification problems. However, they can be extended to solve multi-class classification problems using a variety of strategies. One of the most commonly used strategies is one-versus-all (OvA). Basically, by using OvA, an M-class problem is decomposed into a set of two-class problems. The idea is to train M SVMs where the *i*-th classifier is trained to separate the class *i* from all other classes. The reason for its wide use is the ease of implementation and the speed of the training and testing phase.

The random forest [21] classifier consists of a number of decision trees. Each tree is grown using some form on randomization. Each non-leaf node contains a test that best splits the data space. Classification is performed by sending a sample is down each tree and assigning it the label of the terminal node it ends up in. At the end the average vote of all trees is reported as the result of the classification. Random forest is very efficient with large datasets and high dimensional data.

The k-nearest neighbour classifier [22] is a widely used method for classifying samples based on closest training samples. The process of classification is performed using a voting. The sample is classified by the majority vote of its k-nearest neighbours and k is the parameter that can be adjusted. It specifies the number of samples which will be considered in the classification process. It is among the simplest classifiers in terms of implementation. The main problem is that this method is biased towards the dominant classes in the dataset.

V. Experimental setup

For the investigation performed in the paper, images from the MIAS database [23] were used. This database is a collection of 326 annotated images of normal and abnormal tissue. The database also contains the coordinates and the radius of each abnormality on the image. An example is shown on Fig. 2.

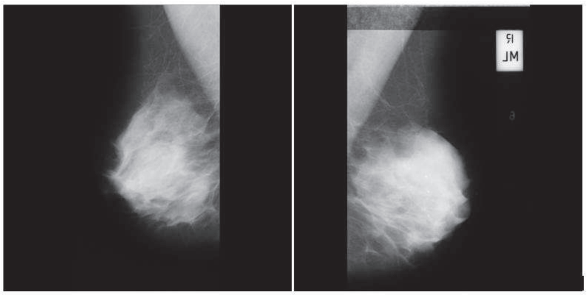

Figure 2. Mammographic image of left and right breast

Two experimental datasets containing the same images were created. The difference between the two datasets is that in the first one the images are split in two classes: normal or abnormal (benign or malignant), and in the second dataset the images are split in three classes: benign, malign and normal.

All feature extraction methods explained earlier were implemented in Matlab 2008 and used in the research. The histogram equalization process was also implemented using Matlab. For the classification methods we used the Weka library [24]. Random forest and k-nearest neighbour are implemented in Weka, while there is only a wrapper for SVM. We used LibSVM library for the SVM classifier [25].

The experiment was conducted in a couple of steps. First, the region of interest was extracted from all images in the dataset. For the abnormal images the region of interest is extracted according to the information given in the dataset. For the normal images a random 50x50 pixels region was extracted from the breast tissue. The second step is applying histogram equalization on the extracted regions of interest. The next step is the feature extraction from the normalized image regions using all feature extraction methods. The last step of the process is performing the classification for every descriptor separately. The testing phase is performed using 10-fold cross validation because of the small number of images in the dataset.

The procedure was performed in the two-class and the three-class dataset using all previous steps. Then the procedure was performed on the same datasets but without the region extraction process. Instead of the extracted region in the second experiment the feature extraction process is performed on the whole image.

The task of the classifier in both experiments is to recognize the abnormality in the tissue and to classify it accordingly.

The goal of the research is to conclude which descriptor is the best for describing mammographic images and which classification method will yield the best results i.e. to find out which combination of descriptor and classifier will give the best results.

VI. Results and discussion

Table 1 depicts the classification accuracy for the two-class problem. The table is divided in two parts. The first half depicts the results when the features are extracted only from the abnormal region. The second half depicts the results when the features are extracted from the whole image. It can be noted than the results are better when the classification is performed over the part of the image rather than the entire image. This is expected since the entire image carries additional information which is irrelevant in the classification process.

The best classification accuracy is in the case of the random forest classifier when the images are described using GLDM descriptor with 99.08% classification accuracy. The best results for KNN and SVM are also in the case of the GLDM descriptor with 97.23% and 98.15% classification accuracy, respectively.

Table 1: Classification accuracy for two-class problem

Descriptor / Classifier	**Classification precision per classifier from regions (%)**			**Classification precision per classifier from image (%)**		
	SVM	***KNN***	***RF***	***SVM***	***KNN***	***RF***
LBP	95,38	94.15	94.46	68.09	61.35	62.58
GLDM	98.15	97.23	99.08	68.09	59.82	59.20
GLRLM	96.61	86.70	90.46	68.09	59	59.51
Haralick feat.	81.54	77.23	86.46	68.09	62.27	57.67
Gabor feat.	84	74	78	68.09	59.51	61
Combined	91.08	84.92	99.08	68.09	56.44	62.88

The results for the three class problem are presented in Table 2. In the case of the three class problem the results are similar to the two class problem. Again, the best classification performance is in the case of the random forest classifier when the images are described using GLDM descriptor with a classification accuracy of 87%.

Table 2: Classification accuracy for three-class problem

Descriptor / Classifier	**Classification precision per classifier from regions (%)**			**Classification precision per classifier from image (%)**		
	SVM	***KNN***	***RF***	***SVM***	***KNN***	***RF***
LBP	82.77	83.08	80.92	67.48	58.28	64.72
GLDM	84	86	87	67.48	59	59
GLRLM	76.31	76	80.62	67.48	58.59	63.19
Haralick feat.	74.15	68.31	74.15	67.48	58.28	57.67
Gabor feat.	71.38	67.08	71.69	65.95	57.06	61.96
Combined	77.54	75.38	85.85	67.48	58.28	62.27

The reason for GLDM's supremacy is the fact, that different displacements were combined in one feature vector. The GLDM implementation used for the purpose of the paper calculates the five features for every displacement (1,2,3,4,5) and then concatenates them in one single feature vector. This means that the GLDM descriptor describes various sizes of textures appearing in the images, thus providing a richer description of the image.

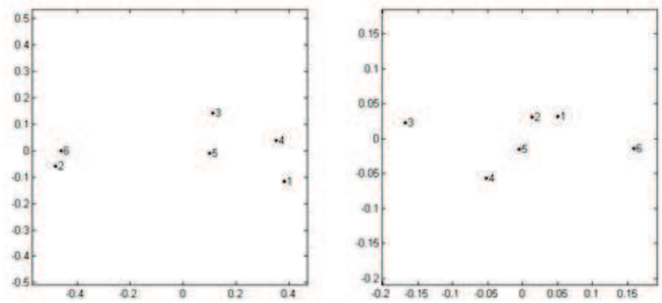

Figure 3. Two-dimensional feature representation using multi-dimensional scaling a) when using image regions b) when using images

If the thoroughly analysis on the results is performed, it can be noted that the combined descriptor provides poorer image description. The authors of [26] propose a method for analysing the correlation between various features. The method uses multi-dimensional scaling to represent the features in two-dimensional space. Figure 3 depicts the results of the multidimensional scaling for our feature in two cases: a) when the features are extracted from image patches/regions and b) when the features are extracted from the whole image. There are six features numbered in the following manner: 1 - Gabor, 2 - GLDM, 3 – GLRLM, 4 – Haralick, 5 – Combined, 6 – LBP. On Figure 3b most of the features that describe similar characteristics are presented, thus they are close in the two-dimensional space. As shown on Figure 3a, the features are more dispersed, but basically the two best performing descriptors are close in the two-dimensional space. This means that they have extracted similar features from the image patches. The combined descriptor is closer to GLRLM, Haralick and Gabor, which provide poorer performance, thus explaining the poorer performance of the combined descriptor.

VII. Conclusion

Automated analysis of mammographic images is a very important and complex challenge. In this paper, five texture descriptors were used to describe the images, or image patches. Then, the resulting descriptors were classified using three classification algorithms. The best results were achieved in the case of the GLDM descriptor when the images were classified using random forest classifier. Overall better results are achieved when image patches are described and classified, rather when the same process is performed over the entire image.

References

[1] O. Mangasarian, W. Street, and W. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.

[2] A. Majid, E. de Paredes, R. Doherty, N. Sharma, and X. Salvador, "Missed breast carcinoma: Pitfalls and pearls1," *Radiographics*, vol. 23, no. 4, p. 881, 2003.

[3] M. Sampat, M. Markey, A. Bovik *et al.*, "Computer-aided detection and diagnosis in mammography," *Handbook of image and video processing*, vol. 10, no. 4, pp. 1195–1217, 2005.

[4] R. Strickland and H. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *Medical Imaging, IEEE Transactions on*, vol. 15, no. 2, pp. 218–229, 1996.

[5] M. Giger, F. Yin, K. Doi, C. Metz, R. Schmidt, and C. Vyborny, "Investigation of methods for the computerized detection and analysis of mammographic masses," in *Proceedings of SPIE*, vol. 1233, 1990, p. 183.

[6] S. Liu and E. J. Delp, "Multiresolution detection of stellate lesions in mammograms," in *In Proceedings of the IEEE International Conference on Image Processing*, 1997, pp. 109–112.

[7] D. Brzakovic, X. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," *Medical Imaging, IEEE Transactions on*, vol. 9, no. 3, pp. 233–241, 1990.

[8] A. Mohanty and S. Lenka, "Efficient image mining technique for classification of mammograms to detect breast cancer," *Special Issue of IJCCT*, vol. 2, no. 2, p. 3.

[9] E. Cheng, N. Xie, H. Ling, P. Bakic, A. Maidment, and V. Megalooikonomou, "Mammographic image classification using histogram intersection," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*. IEEE, 2010, pp. 197–200.

[10] L. Martins, A. Santos, A. Silva, and A. Paiva, "Classification of normal, benign and malignant tissues using co-occurrence matrix and bayesian neural network in mammographic images," in *Neural Networks, 2006. SBRN'06. Ninth Brazilian Symposium on*. IEEE, 2006, pp. 24–29.

[11] J. Weszka, C. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 4, pp. 269–285, 1976.

[12] R. Conners and C. Harlow, "A theoretical comparison of texture algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 3, pp. 204–222, 1980.

[13] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 971–987, 2002.

[15] M. Galloway, "Texture analysis using gray level run lengths," *Computer graphics and image processing*, vol. 4, no. 2, pp. 172–179, 1975.

[16] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 3, no. 6, pp. 610–621, 1973.

[17] D. Zhang, A. Wong, M. Indrawan, and G. Lu, "Content-based image retrieval using gabor texture features," *IEEE Transactions PAMI*, pp. 13–15, 2000.

[18] V. Vapnik, *The nature of statistical learning theory*. Springer Verlag, 2000.

[19] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[20] T. Joachims, "Making large scale svm learning practical," 1999.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[23] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok *et al.*, "The mammographic image analysis society digital mammogram database," 1994.

[24] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.

[25] C. Chang and C. Lin, "Libsvm: a library for support vector machines," 2001

[26] T. Deselaers, D. Keysers, and H. Ney, "Features for Image Retrieval: A Quantitative Comparison" , *DAGM-Symposium*, pp. 228-236, 2004.